

巢湖流域非点源污染物来源的模糊聚类 对应分析方法*

王宗志^{1,2} 金菊良² 洪天求³

(1 南京水利科学研究所, 南京 210029)

(2 合肥工业大学土木建筑工程学院, 合肥 230009)

(3 合肥工业大学资源与环境工程学院, 合肥 230009)

FUZZY CLUSTERING BASED CORRESPONDENCE FACTOR ANALYSIS METHOD FOR NON-POINT SOURCE POLLUTION IN CHAOHU DRAINAGE BASIN

Wang Zongzhi^{1,2} Jin Juliang² Hong Tianqiu³

(1 Nanjing Hydraulic Research Institute, Nanjing 210029, China)

(2 College of Civil Engineering, Hefei University of Technology, Hefei 230009, China)

(3 College of Natural Resources and Environmental Engineering, Hefei University of Technology, Hefei 230009, China)

关键词 巢湖流域; 非点源污染; 对应分析; 模糊 C-均值聚类; 遗传算法

中图分类号 X524 文献标识码 A

对应分析, 又称对应因子分析 (Correspondence Factor Analysis, CFA), 是由法国学者 Jean-Paul Benzecri 在 20 世纪 70 年代提出的一种几何分析技术^[1~4], 是主成分分析技术的拓广^[2]。CFA 依靠主成分分析的降维手段, 使高维问题变为低维问题; 采取独特的数据预处理方式使样本矩阵 Q 和变量矩阵 R 呈对称性, 并从数学上证明了 R 与 Q 具有正交关系, 有相同的特征值, 从而可以在相同标度的因子轴坐标系中同时标度变量和样本^[1~7]。CFA 依据被标注点群的聚散状况, 解释与判断变量之间、样本之间以及变量与样本之间的多重内在联系, 已在众多领域得到了广泛的应用。CFA 的解析原理是^[1,3]: 在同一标度的因子平面或空间里, 临近的变量点, 表示它们密切相关或来自同一源; 临近的样本点, 则表示它们密切相关或属于同一类; 靠近样本点的变量点, 表示这些变量对这些样本点具有造成、导致或其他某种关系, 其相关程度可由它们之间的距离表示。聚类是人类认识事物的重要手段^[8], 因而, 人们在使用 CFA 时往往要将这些点群分成几类, 使类间差异

性和类内相似性都尽量大, 以更方便揭示其多重内在联系。显然, 这是一个聚类数目未知的模糊聚类问题, 若根据经验随意确定, 往往导致不正确的分析结果^[9]。为此, 本研究将 Xie 等^[9,10] 提出的模糊聚类有效性函数 S 引入模糊 C-均值聚类 (Fuzzy C-means Clustering, FCM)^[8~10] 算法中, 取 S 随聚类中心数目 C 变化曲线中最小值对应的 C 作为最佳聚类数目, 并用实码加速遗传算法 (Real coding based Accelerating Genetic Algorithm, RAGA)^[11] 优化该复杂非线性模型, 同时获取最佳聚类数、聚类中心及各点隶属各中心的隶属度, 实现了聚类数目未知下的最佳快速排序, 提出了基于模糊 C-均值聚类的对应分析方法 (Correspondence Factor Analysis based on Fuzzy C-means Clustering, CFA-FCM)。将 CFA-FCM 用于巢湖的非点源污染研究中^[9], 剖析了巢湖流域不同土地利用方式与营养物地表径流输入之间的对应关系。

1 CFA-FCM 的建立

CFA-FCM 的建立包括以下 7 个步骤: (1) 原始数

* 国家重大基础研究前期研究专项“巢湖输入型非点源污染的源解析及控制”(2003CCC00400)、国家自然科学基金项目(50579009, 70471090)、教育部优秀青年教师资助计划、安徽省自然科学基金项目(01045102, 01045409) 共同资助

作者简介: 王宗志(1977~), 男, 山东邹平人, 南京水利科学研究所博士研究生, 从事水资源系统工程研究

收稿日期: 2004-09-15; 收到修改稿日期: 2005-01-12

据预处理^[1,2,5~7]。设原始数据矩阵为 $X = \{x_{ij} | i = 1, 2, \dots, n; j = 1, 2, \dots, m\}$, 其中 n 为样本数目, m 为变量数目。将 X 转化为新数据阵 $Z = \{z_{ij} | i = 1, 2, \dots, n; j = 1, 2, \dots, m\}$:

$$z_{ij} = \frac{x_{ij} - x_i \cdot x_j / T}{(x_i \cdot x_j)^{0.5}} \quad (i = 1, 2, \dots, n, j = 1, 2, \dots, m) \quad (1)$$

式中, $x_i = \sum_{j=1}^m x_{ij}$, $x_j = \sum_{i=1}^n x_{ij}$, $T = \sum_{i=1}^n \sum_{j=1}^m x_{ij}$ 。

(2) 计算变量的协方差阵 $R = Z^T Z = \{r_{jk} | j = 1, 2, \dots, m, k = 1, 2, \dots, m\}$ ^[1,2,5~7]:

$$r_{jk} = \sum_{i=1}^n z_{ij} \times z_{ik} \quad (2)$$

(3) 从 R 阵出发进行 R 型因子分析。先用 Jacobi 法求矩阵 R 的特征值及相应的特征向量, 并按特征值从大到小排序, 不失一般性, 记这些特征值为: $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$, 相应特征向量为 u_1, u_2, \dots, u_m 。然后根据累计方差贡献率 $\sum_{j=1}^p \lambda_j / \sum_{j=1}^m \lambda_j \geq 0.85$, 确定前 p 个为主因子 ($p \geq 2$)。计算 R 型因子载荷阵 A :

$$A = \begin{bmatrix} u_{11}(\lambda_1)^{1/2}, u_{12}(\lambda_2)^{1/2}, \dots, u_{1p}(\lambda_p)^{1/2} \\ u_{21}(\lambda_1)^{1/2}, u_{22}(\lambda_2)^{1/2}, \dots, u_{2p}(\lambda_p)^{1/2} \\ \dots & \dots & \dots \\ u_{m1}(\lambda_1)^{1/2}, u_{m2}(\lambda_2)^{1/2}, \dots, u_{mp}(\lambda_p)^{1/2} \end{bmatrix} \quad (3)$$

并在两两因子轴上作变量点图。

(4) 从 $Q = ZZ^T$ 出发进行 Q 型因子分析^[5,7]。由于 $Ru_j = Z^T Z u_j = \lambda_j u_j$, 两边同时左乘 Z , 得 $ZZ^T Z u_j = Q(Z u_j) = \lambda_j (Z u_j)$ 。因此, Q 与 R 的非零特征值相同, 即 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$ 也是 Q 的特征值, 相应的特征向量为 $V_j = Z u_j (j = 1, 2, \dots, m)$ 。不失一般性, 将 Q 阵的特征向量单位化后仍记为 $V_j (j = 1, 2, \dots, m)$ 。计算 Q 型因子载荷阵为 B :

$$B = \begin{bmatrix} v_{11}(\lambda_1)^{1/2}, v_{12}(\lambda_2)^{1/2}, \dots, v_{1p}(\lambda_p)^{1/2} \\ v_{21}(\lambda_1)^{1/2}, v_{22}(\lambda_2)^{1/2}, \dots, v_{2p}(\lambda_p)^{1/2} \\ \dots & \dots & \dots \\ v_{n1}(\lambda_1)^{1/2}, v_{n2}(\lambda_2)^{1/2}, \dots, v_{np}(\lambda_p)^{1/2} \end{bmatrix} \quad (4)$$

在对应的 R 型因子平面上作样本点图。

(5) 置聚类数目初值 $C = 2$, 用 FCM 算法确定同一因子平面上聚类点 (包括所有变量点和样本点, 记为 $x_i, i = 1, 2, \dots, m + n$) 的聚类中心和各点隶属聚类中心的隶属度。常规的 FCM 算法是通过预先随机给定各聚类点隶属各聚类中心的隶属度、以最小

化准则函数

$$\min f(O) = \sum_{k=1}^C \sum_{i=1}^{m+n} w_{ki}^r |x_i - O_k|^2 \quad (5)$$

迭代求聚类中心和各点隶属各聚类中心的隶属度的, 计算复杂、精度不高。式 (5) 中, x_i 为第 i 个待聚类点, O_k 为第 k 个聚类中心, $|x_i - O_k|$ 为欧氏距离,

$w_{ki} = \frac{\left[\frac{1}{\sum_{k=1}^C |x_i - O_k|^2} \right]^{1/r-1}}{\left[\frac{1}{\sum_{k=1}^C |x_i - O_k|^2} \right]^{1/r-1}}$ 为 i 点隶属聚类中心 O_k 的隶属度, C 为聚类中心数目, $r \in [1.5, 30]$ 为模糊加权指数^[9]。模拟生物优胜劣汰规则与群体内部染色体信息交换机制的 RAGA 是一种通用的全局优化方法, 求解这一非线性问题较为简便有效^[11]。

据此, 考虑到 CFA 中聚类中心变量数目 (聚类中心一般是二维的) 远小于聚类点的数目, 以聚类中心和隶属度为优化变量、以式 (5) 为优化准则函数用 RAGA 可快速求解该问题, 当优化准则函数值不再减小时, 终止算法的运行, 求得聚类中心 $O_k (k = 1, 2, \dots, C)$ 、各点隶属各聚类中心的隶属度 $w_{ki} (k = 1, 2, \dots, C; i = 1, 2, \dots, m + n)$ 和最优化准则函数值, 记为 J_m 。

(6) 计算模糊聚类有效性函数 S 值^[9,10]:

$$S = \text{Comp} / \text{Sep} \quad (6)$$

式中, $\text{Comp} = J_m / (m + n)$ 为模糊聚类紧致性函数; $\text{Sep} = \min_{i,k} |O_i - O_k|^2$ 为模糊聚类分离性函数。显然, S 越小, 分类越有效^[10]。

(7) 置 $C = C + 1$, 重复步骤 5, 步骤 6, 直至 $C = \min(m, n)$ 。取 S 随 C 变化曲线中最小的 S 值对应的 C 作为最佳聚类数目, 并按最大隶属原则将待聚类点 $\{x_{ij}\}$ 划分为 C 类。

2 实例分析

以巢湖流域西部塘西地区为研究区域, 剖析不同土地利用方式与营养物地表径流输入之间的内在联系, 为巢湖流域非点源污染防治提供依据, 有关监测数据见表 1^[7]。

这是由 6 个变量 (A ~ F) 和 11 个样本 (土地类型) 组成的原始数据阵, 即 $X = \{x_{ij} | i = 1, 2, \dots, 11; j = 1, 2, \dots, 6\}$, X 经式 (1) 和式 (2) 得矩阵 R , 由 Jacobi 法求得矩阵 R 的特征值 $\lambda_1 = 0.1901$, $\lambda_2 = 0.018$, $\lambda_3 = 0.018$, $\lambda_4 = 0.018$, $\lambda_5 = 0.018$, $\lambda_6 = 0.018$, λ_1 的方差贡献率为 86.6%; $\lambda_1 + \lambda_2$ 的累计方差贡献率为 94.8%, 这表明前两个主因子可代表变量或样本的全部信息的 94.8%。

表 1 巢湖流域塘西地区不同土地利用类型地表径流检测数据^[7] (mg L⁻¹)

土地类型	总磷(A)	总氮(B)	总 COD(C)	水溶性磷(D)	水溶性氮(E)	水溶性 COD(F)
大豆地(1)	1.14	12.78	15.06	0.53	4.94	5.68
冲畝水田(2)	0.08	2.47	10.12	0.03	1.33	5.80
菜园地(3)	0.41	7.62	15.11	0.26	5.75	7.32
山芋地(4)	0.60	4.36	51.48	0.07	3.01	5.76
小麦地(5)	0.38	5.94	33.45	0.15	4.46	5.11
集镇道路(6)	0.56	7.01	184.6	0.45	4.31	5.15
山坡地(7)	0.22	5.04	10.90	0.11	4.42	9.95
饲养地(8)	0.25	8.53	48.11	0.14	7.79	5.07
农村道路(9)	0.20	6.02	6.29	0.07	5.01	5.67
湖滩苇地(10)	0.41	5.30	111.7	0.71	11.54	8.39
荒地(11)	0.11	2.32	8.40	0.11	1.22	2.89

R 型和 Q 型因子载荷阵见表 2。以第 1 主成分载荷(F₁)为横坐标,第 2 主成分载荷(F₂)为纵坐标,将各载荷点描绘在同一因子平面上,见图 1。取模

糊加权指数 $r=2$, $C [2,6]$,运行 CFA-FCM 建模步骤(5)、步骤(6)和步骤(7),得到的有效性函数 S 随聚类数目 C 变化的曲线见图 2。

表 2 R 和 Q 因子载荷阵

载荷点	F ₁	F ₂	载荷点	F ₁	F ₂
A	0.053 4	- 0.043 9	4	- 0.044 7	0.011 1
B	0.251 1	- 0.079 2	5	0.027 9	- 0.003 2
C	- 0.228 4	- 0.002 7	6	- 0.230 0	- 0.020 8
D	0.025 9	- 0.023 4	7	0.136 3	0.032 3
E	0.159 9	0.017 5	8	0.022 6	- 0.019 7
F	0.213 9	0.095 1	9	0.184 6	0.018 6
1	0.193 7	- 0.095 1	10	- 0.094 6	0.022 7
2	0.082 3	0.074 7	11	0.055 5	0.018 3
3	0.157 9	0.011 6			

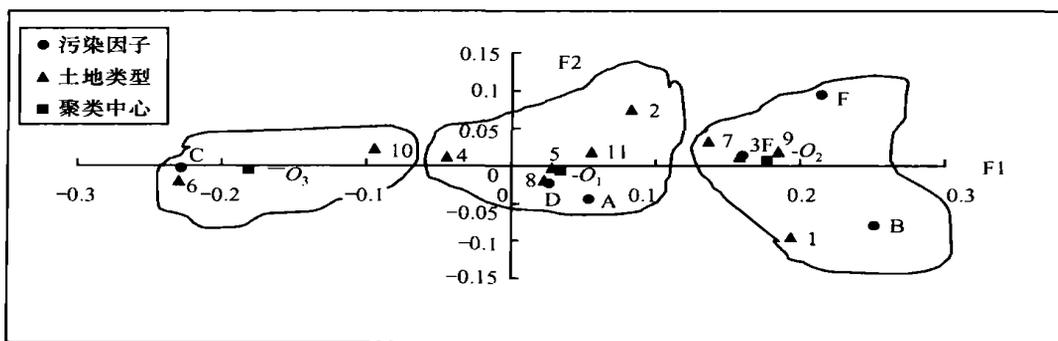


图 1 土地类型、污染因子及其聚类中心在因子平面上的投影

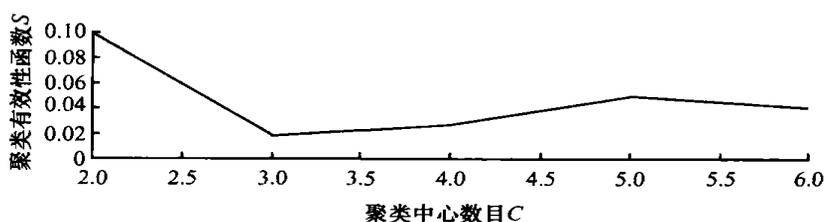


图 2 聚类有效性函数 S 与聚类中心数目 C 的关系

根据图 2 可得该点群的最佳分类数目 $C = 3$ 。分别用常规的迭代法^[9,10]和 RAGA 求解,其聚类中心见表 3,载荷点属于各聚类中心的隶属度见表 4。由表 3 知,RAGA 的最小准则函数值小于迭代法最小准则函数值,说明 RAGA 更有效。

根据表 4 和最大隶属度原则,可将聚类点分成如下三类:(A,D,2,4,5,8,11),(B,E,F,1,3,7,9),(C,6,10)。这说明巢湖流域的土地利用方式与地表径流营养盐的输入之间存在明显的对应关系:(1)总磷和水溶性磷的污染主要来自冲

刷水田、山芋地、小麦地、饲养地、荒地等的径流输入;(2)总氮、水溶性氮和水溶性 COD 的输入主要来自大豆地、菜园地、山坡地、农村道路等的径流输入;(3)总有机物主要来自集镇道路和湖滩苇地的径流输入;(4)变量总磷和水溶性磷属于同一源;(5)变量总氮、水溶性氮和水溶性 COD 属于同一源;(6)总 COD 的成因单独成一类。这些结论将有利于指导我们合理布置巢湖流域的土地利用方式和作物的种植结构,以最有效地控制流域非点源污染。

表 3 用迭代法与 RAGA 求得的聚类中心坐标及最优化准则函数值

算法	聚类中心	F_1	F_2	J_m
迭代法	O_1	0.030 1	- 0.005 8	0.051
	O_2	0.177 3	0.006 4	
	O_3	- 0.211 7	- 0.007 9	
RAGA	O_1	0.033 8	- 0.006 8	0.049
	O_2	0.176 7	0.006 7	
	O_3	- 0.181 9	- 0.005 2	

表 4 载荷点属于各聚类中心的隶属度

算法	聚类中心	A	B	C	D	E	F	1	2	3	4	5	6	7	8	9	10	11
迭代法	O_1	0.878	0.182	0.005	0.981	0.019	0.167	0.222	0.564	0.024	0.753	0.999	0.007	0.154	0.986	0.008	0.427	0.911
	O_2	0.098	0.773	0.001	0.013	0.979	0.795	0.732	0.380	0.973	0.090	0.001	0.003	0.830	0.010	0.991	0.094	0.074
	O_3	0.024	0.045	0.994	0.006	0.002	0.038	0.046	0.056	0.003	0.157	0.000	0.990	0.016	0.003	0.001	0.479	0.015
RAGA	O_1	0.885	0.187	0.030	0.979	0.018	0.169	0.230	0.561	0.023	0.680	0.997	0.035	0.157	0.981	0.009	0.303	0.915
	O_2	0.088	0.762	0.013	0.013	0.979	0.788	0.719	0.373	0.973	0.090	0.002	0.015	0.825	0.012	0.990	0.071	0.067
	O_3	0.027	0.051	0.957	0.008	0.003	0.043	0.051	0.066	0.003	0.230	0.001	0.950	0.018	0.007	0.001	0.626	0.018

3 结 论

根据流域非点源污染物输入特征与土地利用方式之间的对应关系,提出了基于模糊 C-均值聚类的对应分析模型(CFA-FCM)。CFA-FCM 的应用结果表明:巢湖流域土地利用方式与营养物径流输入特征之间存在明显的 3 类对应关系;CFA-FCM 建模方便、分类合理,在流域非点源污染源解析、不同树种生长对土壤肥力的影响分析^[12]、流域化肥适宜用量对地下水的影响分析^[13]等中也具有应用价值。

参 考 文 献

- [1] 卢崇飞,高惠璇,叶文虎. 环境数理统计原理及应用程序. 北京:高等教育出版社,1989
- [2] 陈上珠,张奕,徐民. 多元对应分析在抽样调查分析中的应用. 数理统计与管理,2001,20(4): 17~24
- [3] 艾云灿,孟繁梅,高培基,等. 尝试利用 R-Q 对应因子分析识别分群融合子代. 生物工程学报,1996,12(1): 28~33

- [4] 王晓鹏. 水质评价的对应分析方法及在湟水流域的应用. 青海师范大学学报(自然科学版),2001,(1): 66~70
- [5] Tao F M, Xia L X. Stepwise correspondence analysis and study of physical characteristics of Chinese ethnic groups. Journal of Biomathematics, 2003, 18(2): 139~145
- [6] 王学仁. 地质数据的多变量统计分析. 北京:科学出版社,1986
- [7] 阎伍玖,王心源. 巢湖流域非点源污染初步研究. 地理科学,1998, 18(3):263~267
- [8] 凌文燕,李德毅. 基于核密度估计的层次聚类算法. 系统仿真学报,2004, 16(2): 302~306
- [9] 熊浩,周杰,丁萃菁. 模块化模糊网络在电站锅炉故障诊断中的应用研究. 计算机仿真,2004, 21(2):102~104
- [10] Xie XL, Beni G. A new fuzzy clustering validity and its application to color image segmentation. Proceedings of the 1991 IEEE International Symposium on Intelligent Control, Arlington, Virginia, U.S.A, 1991
- [11] 金菊良,丁晶. 水资源系统工程. 成都:四川科学技术出版社,2002
- [12] 薛立,赖日石,陈红跃,等. 不同阔叶树种的生长及其对赤红壤肥力的影响. 土壤学报,2003, 40(5):795~799
- [13] 王德建,林静慧,孙瑞娟,等. 太湖地区稻麦高产的氮肥适宜用量及其对地下水的影响. 土壤学报,2003, 40(3):426~432